DOCUMENT RESUME

ED 178 460 SP 011 349

AUTHOR Gage, N. L.
TITLE Four Cheers for Research on Teaching.
PUB DATE [78]
NOTE 44p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Educational Research; *Effective Teaching; *Evaluation Methods; *Reliability; *Research Criteria; *Research Methodology; Research Problems; Student Attitudes; Teacher Behavior

ABSTRACT

This paper reviews the reliability of several efforts by educational researchers to determine variables in teacher behavior that influence student attitudes and achievement. In particular, works by Rosenshine (1971), Dunkin and Biddle (1974), and Rosenshine and Furst (1970) are examined and evaluated in terms of research criteria developed by Heath and Nielson (1974) and Glass (1976). Efforts to replicate the results of educational research are cited, and the significance of various studies are discussed. The appendices include tables illustrating the results of field studies concerning the effect of teacher indirectness on pupil achievement as well as an extensive bibliography. (IH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
N. L. Gage
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED178460

Four Cheers for Research on Teaching[1]

N. L. Gage

Stanford University

E. M. Forster's Two Cheers for Democracy betokened a cautious and skeptical optimism. Three cheers stand for congratulations on a job well done. By four cheers, I mean support and encouragement, a way of saying, "Keep it up! Try harder! You'll make it yet!"

For the field of research on teaching has not made it yet. It has won adherents, but not the substantial victories that research workers dream of. We have not yet achieved our polio vaccine or double helix.

Science depends heavily on replication for the establishment of valid knowledge. The inability of scientists to confirm a finding is taken as evidence of its invalidity. When scientists insist upon a finding that no one else can confirm, they begin to impugn their own integrity.

The more significant a study and its results, the more it attracts attempts at replication. Thus, the Rosenthal-Jacobson findings of increased pupil IQ resulting from teachers' being misled to expect such increases elicited a host of replication efforts. (According to Baker and Crist, 1971, they resulted in no confirmations.) The "Coleman Report" similarly led to much reanalysis and replication (Do Teachers Make a Difference, 1970; Mosteller & Moynihan, 1972; Averch, et al., 1971).

Far more than the statistical significance of any single study, confirmation by independent studies is relied upon by behavioral scientists before they begin to take a finding seriously. Reasons for this tendency

SP011349

are readily apparent. Statistical significance in itself merely tells us that, under the particular conditions of a given study, a result relatively nonattributable to chance was obtained. The particular conditions of that result and whether it would recur under altered conditions (where the alterations may be small or large) are matters of crucial importance. Similarly, the biases, usually unconscious, of the single investigator or team often tend to make the results of their studies turn out the way they want them to (see the concept of Commitment in Dunkin and Biddle, 1974, pp. 51-52).

One major safeguard against the occurrence of statistically significant results because of special but unrecognized conditions and investigators' commitments is to carry out a similar study, under similar conditions, with similar variables, using comparable methods. In short, what we want in most fields of research, before we become truly impressed, is replication.

It is also largely for these reasons that reviews of research become important. Reviews bring together the results of independent studies and provide assessments of the degrees to which the individual studies are valid and have yielded cumulative results worthy of scientific respect.

The Bad Press

On the basis of such reviews, research on teaching has had a bad press --a long history of assertions that the field was not getting anywhere very fast. I wrote one of the first of these in 1952, as the Young Turk on the AERA Committee on the Criteria of Teacher Effectiveness (1952). The Committee report bemoaned the lack of progress in research on teaching during the first half of this century. Then, in the Handbook of Research on Teaching, similar statements were made by the authors of

the chapters on teaching methods (Wallen & Travers, 1963), teacher personality and characteristics (Getzels & Jackson, 1963), and social interaction in the classroom (Withall & Lewis, 1963). In 1966, the Coleman Report alleged that teacher variables accounted for only tiny proportions of the variance in student achievement. When Stephens (1967) reviewed the reviews, he was impelled to a theory of spontaneous schooling --teachers will do just as well as they can if they do what comes naturally, because none of the attempts to improve on spontaneous behavior had improved school learning. Then the Coleman conclusions were supported by the re-analyses contained in the Mosteller-Moynihan volume, On Equality of Educational Opportunity (Mosteller & Moynihan, eds., 1972). When Averch and his colleagues (1972) reviewed the literature, they also found against the efficacy of variables in teachers and teaching for making a difference in what students achieved.

Much of this writing has been misinterpreted as signifying that teaching makes little or no difference. But the studies have not compared teaching with no teaching. Rather they have compared different kinds of teachers and teaching and have concluded, essentially, that one kind is about as effective as another. In short, teaching variables make little difference.

This kind of conclusion has often been based on research in which teaching variables never had much of a chance. Thus in the Coleman-type studies, the school rather than the teacher was the unit of analysis, the teacher variables were characteristics rather than types of behavior, and the dependent variables were scores on standardized tests not closely related to what the teachers had been trying to help their pupils learn. Nonetheless, not all of the negative results can be explained in this way.

Much proper research did use the teacher as the unit of analysis, focused on teacher behavior, and used instruction-relevant achievement measures. In short, despite the frequently lugubrious conclusions, a whole generation of investigators has persisted in looking for the variables in teaching behavior that do make some difference in student achievement and attitude.

The present paper is concerned with reviews of that kind of research on teaching. Two volumes containing such reviews have appeared in recent years (Rosenshine, 1971; Dunkin and Biddle, 1974). Apart from the difference in publication dates, and the consequent greater inclusiveness of the Dunkin-Biddle volume, the two books differ in four other respects. First, the Rosenshine volume reviewed studies in which teacher behavior was either rated or counted, while the Dunkin-Biddle volume excluded rating studies. Second, the Rosenshine volume included studies done in either classrooms or laboratories, including simulated classrooms, while the Dunkin-Biddle volume dealt only with studies conducted in classrooms. Third, the Rosenshine volume dealt only with achievement of cognitive objectives, while the Dunkin-Biddle volume dealt also with attitudinal and process (student-behavior-in-the-classroom) outcomes. Fourth, the Rosenshine volume dealt only with the relationships between teacher behavior and student achievement (i.e., process-product relationships), while the Dunkin-Biddle volume dealt also with findings concerning process-occurrence and context-process, presage-process, and process-process relationships.

When Rosenshine and Furst (1970) and Rosenshine (1971) brought the evidence together, it was possible to discern some consistency in the findings of clusters of studies that dealt with a given variable in teacher behavior. But that hope-giving compilation was not to go unchallenged. Soon Heath and Nielson (1974) published what they intended

as a devastating attack on the whole body of literature that Rosenshine and Furst had cited. They found that literature only spottily satisfactory on a list of 13 criteria. The studies cited were characterized by "sterile operational definitions of both teaching and achievement" and "fundamentally weak research designs" (p. 481).

Similar characterizations of this body of research continue to be written. Thus, the research integration efforts have "with remarkable regularity, failed to support the existence of stable and consistent relationships between teacher variables and effectiveness criteria" (Doyle, 1975, p. 2).

The criteria that Heath and Nielson applied seem desirable but not indispensable. Those authors imposed their criteria with altogether unrelenting stringency. Such attempts to improve the rigor of research are, in one sense, altogether laudable, and future research workers ought to pay heed to the criteria that Heath and Nielson urged. But some of their criteria are tantamount to saying, "You can't get there from here." Thus one wonders why they are concerned that each study contain at least 15 teachers and each class at least 15 pupils. Or why they insist that teachers should be randomly assigned to pupil groups. That requirement, violating the realities of school organization, would mean severely curtailing research on teaching in the public schools, for only a rare school administration will permit such an interference. And, in any case, random assignment would greatly reduce the external validity of the research.

Heath and Nielson seem to approach research methodology in quite a different spirit from that of other writers, such as Campbell and Stanley (1963), who wrote that:

> ...this chapter is not intended to substitute a dogma of <u>the</u> 13 acceptable designs for an earlier dogma of <u>the</u> one or <u>the</u> two

> acceptable. Rather, it should encourage an open-minded and exploratory orientation to novel data-collection arrangements... Ten designs have been classified as quasi-experiments lacking optimal control but worth undertaking where better designs are impossible (p. 71).

Similarly, contrast Heath and Nielson with Glass (1976), who wrote that:

> A common method of integrating several studies is to carp on the design and analysis deficiencies of all but a few studies...and then advance the one or two "acceptable" studies as the truth of the matter. This approach takes design and analysis much too seriously...a study with a half dozen design and analysis flaws may still be valid (p. 4).

Thus, Heath and Nielson have given the impression that the 42 studies they examined yielded no useful information. Errors of Type I, in which true null hypotheses are rejected as false, are of course undesirable. But so are errors of Type II, in which false null hypotheses are accepted as true. With their wholesale rejection of the findings of this body of research, Heath and Nielson may have made a gigantic error of Type II.

The Dunkin-Biddle Review

Is there evidence that replicated process-product research on teaching has yielded findings suggesting that variance in teacher behaviors is associated with variance in student achievement and attitude? To attempt an answer, I use the review by Dunkin and Biddle (1974), who listed all of the studies they could find on a given kind of teacher behavior. Some of the kinds of teacher behavior with which they dealt were "indirectness," "praise," "acceptance of pupils' ideas," and "criticism." Their longest list cited 104 studies of teacher warmth and directiveness. Then they

briefly gave the major facts about each study: subjects, methods, and findings. They presented the findings verbally with such statements as the following:

> Teacher "indirectness" is unrelated to pupil achievement [15 correlational studies cited].
>
> In contradiction, it is also found that
>
> Higher teacher "indirectness" is associated with greater pupil achievement [10 correlational studies cited] (p. 115).

Similarly, for experiments, they present such conclusions as the following:

> Teacher "indirectness" is unrelated to achievement of average pupils [5 experiments cited] (p. 116).

Thus, the Dunkin-Biddle description of the findings of these studies is dotted with statements that such-and-such a kind of teacher behavior is (and is not) associated with greater pupil achievement; is (and is not) associated with more positive pupil attitudes. These statements give the by-now-familiar impression that research on teaching yields almost completely equivocal, nonsignificant, and inconsistent findings. To say that a given kind of teacher behavior "is (and is not)" associated with higher or lower pupil achievement, attitude, motivation, fear of failure, self-concept, dependency, or anxiety is tantamount to stating that we can learn just about nothing from the research on that behavior.

How did Dunkin and Biddle arrive at these conclusions? They cited no quantitative results, statistical measures, or levels of significance. They did not describe their method in explicit detail, but referred to " clinical judgments" (p. 453). Upon examination of the studies they cite, it appears that their "clinical judgment" is highly correlated with the statistical significance of individual studies. When one looks into the original studies for the quantitative results omitted by Dunkin and Biddle,

it becomes clear that when Dunkin and Biddle say that teacher behavior X is related to student achievement Y, they do so because the original author (or Dunkin and Biddle) found the result to be statistically significant. In short, Dunkin and Biddle by and large based their interpretation of single studies on the statistical significance of the results of those single studies. And in doing so, they have risked committing, more frequently than not, errors of Type II.

On what reasoning can this charge be based? First, it seems safe to assume that the true relationship, if any exists--between (a) any single variable of teacher behavior and (b) pupil achievement or attitude--is probably low. It would be surprising, indeed, if any single variable, such as teacher "indirectness," by itself accounted for any very substantial portion of the variance in student achievement or attitude. On the face of it, the teaching-learning process is so complex that any single variable of teacher behavior should have only a low correlation (ranging from about $\pm$ .1 to about $\pm$ .4) with student achievement or attitude.

Second, the types of studies of teaching with which we are concerned are typically based on relatively small numbers of teachers. Thus, in 19 studies of the correlation between teacher indirectness and student achievement, as brought together by Dunkin and Biddle (1974, p. 115), the number of teachers involved ranged from 8 to 70, with a median of 15. Only five of the 19 studies used samples with more than 18 teachers.

Third, for a sample of the median size, namely, 15 teachers, it is necessary that a correlation coefficient equal .51 if it is to be significant at the .05 level. The coefficient must equal .64 if it is to be significant at the .01 level.

It follows, in view of these sample sizes and our expectations

concerning the magnitude of the relationships between single variables of teacher behavior and student achievement or attitude, that most of the studies should not be expected to attain statistical significance at the .05 level. To insist upon such significance as a basis for a statement that teacher behavior X is related to pupil outcome Y is almost tantamount to insisting on risking an error of Type II.

Vote Counting

What can be done in a field of research where sample sizes are small and expected relationships are low? Should the available studies be essentially disregarded, as is the case when their findings are interpreted as indicating that X is (and is not) related to Y? Or is it possible to salvage significant and valuable knowledge from these studies?

One attempt at an affirmative answer is "vote counting," considered by Light and Smith (1971). Here each individual study is regarded as having cast a vote. The numbers of significant positive, significant negative, and non-significant relationships are counted. The category with the highest number of studies "is declared the winner" (p. 433).

But taking the statistical significance of single studies into account in this way still runs into the already mentioned problems of process-product studies in teaching--the problems of small Ns and low expected rs. It might seem more reasonable in such vote counting to disregard the statistical significance of single studies, as was done by McKeachie and Kulik (1975) in comparing lecture and discussion methods of college teaching. Going beyond Dubin and Taveggia (1968), who considered only the criterion of mean score on the final examination, McKeachie and Kulik considered three criteria--"factual examination," "retention and higher level thinking," and "attitudes and motivation." For each study, they tallied whether the lecture

method or the discussion method was superior. For only a few of the tallies were they able to indicate that the difference between methods was significant at the .05 level or better. "All other results indicate the actual direction of difference in the experiment" (p. 182), regardless of its statistical significance. Using such vote counting, McKeachie and Kulik found impressively consistent superiority of the discussion method for promoting retention, higher level thinking, attitudes, and motivation, and of the lecture method for promoting achievement on a factual examination. In 21 comparisons on the "factual examination," the lecture method was superior in 12, about equal to discussion methods in 4, and inferior to the discussion method in 5. In seven comparisons on "retention and higher level thinking," the discussion method was superior in all 7. In nine comparisons on efficacy in fostering desirable "attitudes and motivation," the discussion method was superior in 7, tied in 1, and inferior in 1.

Dunkin and Biddle (1974) considered this approach, i.e., reviewing descriptive statistics for their sign alone and not their size. They concluded that this method

> is valid only to the extent that all data pertaining to a given relationship are published. But we already know that some nonsignificant findings are not published. Moreover, the tendency to avoid publication should be greatest when the real-world relationship is weak. Thus to conduct such a nose-count is to tabulate from a biased sample, and possibly from a sample generated by Committed investigators. Use of this technique, then, may generate fictitious findings that are not present in the real world (pp. 452-453).

Their argument that nonsignificant findings tend to remain unpublished is unsupported. The plethora of nonsignificant findings in the literature

of research on teaching calls into question any such argument. Furthermore, the possibility has been investigated.

> Bozarth and Roberts [1972] in a survey of a thousand research articles found that 94% of those that used statistical tests reported significance at the .05 level or less. This was a follow-up on a similar survey done by Sterling [1959] who obtained a comparable figure of 97% (Hunt, 1975, p. 592).

But when Hunt (1975) counted not articles but tests of significance, inasmuch as "virtually no articles appear which contain a single test of significance" (p. 592), he found that, in the first five articles in a recent issue of the Journal of Experimental Psychology, three articles reported tests of significance. Of the 110 tests in these three articles, only 32 or 29% were significant at the .05 level or better. "...it cannot be said that only significant findings are published" (p. 592). That published tests of significance are significant more than 5% of the time at the .05 level should simply indicate that research workers tend to seek and test relationships that are suggested by good scientific reasoning.

In any case, the suspicion that nonsignificant results go unpublished jeopardizes the review of cumulated findings on any relationship in the behavioral sciences. No matter how many positive findings were published and summarized, reviewers entertaining this suspicion would always hold that even more nonsignificant findings have gone unpublished. Hence conclusions based on published findings would forever remain untrustworthy.

The Chi-Square Model

There is an alternative to vote counting, one which takes into account the size of the sample and the magnitude of the relationship or difference.

It is also an alternative to the "cluster approach" advocated by Light and Smith (1971)--an approach which rests on the requirement, usually impossible to meet, that the reviewer have "access to the original data" (p. 448). It is also an alternative to the approach used by Glass (1976), which is applicable only to experiments that yield differences between the means of control and experimental groups. This alternative was described by Jones and Fiske (1953) and attributed by them to Karl and Egon Pearson. It is a method for testing the significance of combined results.

> This model is used to test the hypothesis that the composite p value for the several findings could have occurred by chance. It utilizes the numerical p values instead of classifying them as above or below a specified critical level of significance. This model also assumes the statistical independence of the several results being combined.
>
> The model is based on the proof that any p value can be transformed to a chi-square value with two degrees of freedom, and that the sum of independent chi-squares is distributed as chi-square. The transformation equation is
>
> $$\chi^2 = -2 \log_e p \ldots$$
>
> The composite χ^2 is given by the formula
>
> $$\chi^2 = -2 \sum_{i=1}^{\kappa} \log_e p_i$$
>
> with 2κ degrees of freedom where κ is the number of independent probability values to be combined.... The product of the κ separate p values is the joint probability of the κ independent findings....

As Jones and Fiske pointed out, this model requires (a) a knowledge of the exact probability value associated with each result to be combined, (b) statistical independence of the various results being combined, as is true when each result is obtained from a different sample of individuals, (c) the possibility of considering the studies being combined to belong together on some rational basis, (d) inclusion of all obtained results in the set being analyzed, rather than merely the several promising or nearly significant findings, and (e) the recognition that the test of combined significance bears upon a hypothesis about the cluster of studies as a whole and not one about the individual studies within the cluster.

With the assistance of Dale Schunk, I have applied the chi-square model to four clusters of studies in research on teaching. In each case, the clustering had originally been done by Dunkin and Biddle (1974), and I have accepted that clustering as the basis for the present argument. We obtained the original reports of the studies clustered. From those reports, Schunk extracted the values of the statistical measures (such as correlation coefficients, F ratios, t-ratios, chi-squares, Mann-Whitney Us) and determined the one-tailed exact probability of the value of that statistic. He then converted these probabilities to chi-square values, using the tables developed by Gordon, Loveland, and Cureton (1952). The results of one of these tests of combined significance are shown in Tables 1 and 2.

[Insert Tables 1 and 2 about here]

Table 1 provides information concerning 18 correlational studies of the relationship between teacher indirectness and student achievement, as collated by Dunkin and Biddle (1974, p. 115, paragraphs 5.2-26 and 5.2-26a). The test of the significance of combined results requires

that each result being combined be statistically independent. This requirement means that only one set of results for any single sample of individuals is usable in the test of the significance of combined results. When a single study presents more than one set of results, because more than one dependent variable was used, a choice is necessary. For each study that yielded more than one result, we solved this problem by using, in turn, (a) the most significant result, (b) the least significant result and (c) the mean of the two or more results.

Applied to Table 1, this procedure yielded the results shown in Table 2. It is evident that even the most conservative combination of results, the one that uses the lowest value of chi-square from any study that yielded more than one chi-square value, yields a total chi-square equal to 86.38, which is significant beyond the .001 level for 36 degrees of freedom. Thus, considering the results of these studies as a whole, we can conclude that there is a highly significant relationship between teacher indirectness and student achievement.

For the relationship between teacher praise and pupil attitudes as brought together by Dunkin and Biddle (1974, p. 122), the combined result of five correlational studies was significant at better than the .10 level. When eight studies relating teacher acceptance of pupil ideas to pupil achievement were brought together (Dunkin & Biddle, 1974, p. 124), their combined results were significant, when most conservatively estimated, at the .01 level. For 17 studies of teacher criticism and

disapproval in relation to student achievement (Dunkin & Biddle, 1974, p. 126), the sum of the least significant chi-squares was significant at less than the .10 level, while the sum of the mean chi-squares was significant at less than the .001 level.[2]

These methods and findings indicate that research on teaching has yielded some significant results. These results can be detected when the significance of the combined results of clusters of studies of similar variables is determined. Because individual studies are typically based on very small numbers of teachers, and because the relationship to be expected between any single teacher behavior dimension and student achievement or attitude is low, only the combination of results across studies can provide any valid estimate of the significance, or departure from what can be expected by chance, of the relationships under investigation.

In short, some of the pessimism resulting from previous reviews may now be dispelled. Those who have been concluding that significant relationships have not been found should consider this kind of evidence of significant relationships between teachers' behaviors and student achievement or attitudes.

These tests of combined results do indeed reveal some significant relationships between types of teacher behavior and student achievement or attitude. Thus, the results suggest that seeking such process-product relationships is not altogether fruitless. They bear out the widely held and hard to surrender intuition that how teachers behave makes a difference in what students learn.

Further Considerations

The case for research on teaching according to the process-product model cannot rest on the demonstration that clusters of studies, when their results are combined, have yielded relationships that are statistically significant. We must also be concerned with the possibility that additional variables may vitiate or otherwise modify any given relationship. And we must also consider the magnitude of the differences in outcomes associated with teacher behavior variables.

Interaction Effects

The clusters of studies that turn out to have yielded significant results were performed in several different grade levels, and thus their findings have a kind of generality across grade levels. Further, the studies were done with many different subject matters, such as language and number skills, social studies, science, reading, spelling, and arithmetic. Such variety suggests considerable generality across subject matters.

Thus, we have evidence of relationships of teaching variables to outcomes of a kind about which some educational psychologists have expressed strong doubts. Glass (1972) held that educational research "has hardly produced even tendency statements about tendency statements, i.e., that one has observed a tendency for A to tend to be associated with B sometimes and under some conditions" (p. 10). Glass went on to state that the failure to reveal important, reliable, and replicable relationships--a failure that I have questioned--may be attributable to the fact that

> interactions predominate in the accumulated wisdom of educational research. A relationship among variables appears and disappears so often as a function of extraneous conditions that one genuinely never knows when to expect to see the relationship (p. 13).

Cronbach (1975) found similar inconsistencies in the literature which he traced to the possibility of higher order interactions. Thus, even if an aptitude-treatment interaction is found--such that one treatment is better for students of one kind while another treatment is better for students at a different level of a given aptitude or characteristic--that interaction may itself interact with a third variable.

> Once we attend to interactions, we enter a hall of mirrors that extend to infinity. However far we carry our analysis--to third order or fifth order or any other--untested interactions of a still higher order can be envisioned (p. 119).

But the behavioral sciences, including research on teaching, are no worse off in this respect than the natural sciences. In physics, for example, we have relationships about as firm as anyone could wish for. But, considered one at a time, these relationships hold only under laboratory conditions and not in real life. When it comes to real-life phenomena, physical scientists must apply many of those laws in combination. Only one example need be cited: what might be called the interactions that complicate the motion of an object released from the mast of a moving ship. First, the motion can be resolved into horizontal and vertical components, but this

> is only the first step in an essentially infinite chain of resolutions. If one wants more detail about the motion, other laws enter.

> The appearance of the Coriolis force is responsible for an eastward deflection of the object. The laws for falling bodies in real media at various Reynolds numbers have to enter to calculate the effect of friction and turbulence. The more detail one wants to know, the more resolutions become necessary (Holton, 1975, p. 330).

Thus, even simple laws of mechanics interact with a host of other laws. But physical scientists do not denigrate their main effects simply because interaction effects also occur. They simply take more laws into account.

Similarly, the relationships that may be found between variables in teacher behavior and what students learn may need to be applied with due regard for other factors, such as the ability and anxiety level of the pupils, the objectives, the characteristics of other pupils in the class, or the specific task to be taught at any given moment. Until those interacting factors are shown to be important, we can use whatever main effects come to hand. The generalizations that teachers and other appliers of general relationships work with must be combined in complex ways as adaptations are made to the particulars of a specific application.

The Importance of Low Correlations

Despite the significance of the combined results of clusters of studies, the average correlation in those clusters seems low. How useful can low correlations be? If the relationship between teacher acceptance of student ideas and student achievement is only about .3, can that relationship be used as the basis for improving what a teacher does and what her students learn? To gain perspective, we can consider another weak relationship--that between tobacco smoking and lung cancer--which is reflected in a phi-coefficient of about .14, as shown in Table 3

(based on data presented in Lilienfeld, Pedersen, & Dowd, 1967, p. 78). That coefficient is thus of about the same magnitude as that which might be found for any single variable of teacher behavior in relation to pupil achievement. And, of course, the correlation of .14 between smoking and lung cancer has been made the basis for a massive public health movement, Federal legislation concerning the advertising of cigarettes, and changes in the extremely strong habits of millions of persons. Relationships need not be strong to be important.

The teacher's behaving more frequently in one way rather than another costs very little. The effects on student achievement, when multiplied by the number of students affected by the teacher over a number of years, can become great.

The Additivity of Effects

And, if more than one such teacher behavior is changed, it is not inconceivable that the total effect, discernible in a multiple correlation or the sum of the main effects of a number of independent teacher behavior variables, could become substantial.

Thus far, most studies of relationships between teacher behavior and student achievement have been studies of zero-order correlations. Yet a few multiple correlations can be cited that promise tantalizingly impressive total contributions of teacher behavior. For 17 teachers, Wright and Nuthall (1970) obtained a multiple correlation of .89 between six teacher behaviors and pupil achievement. Stallings and Kaskowitz (1974, p. 316) reported that contributions of teacher variables changed zero-order *r*s from .71 and .65 (for the initial Wide Range Achievement Test) to multiple *R*s of .86 and .89 with reading achievement in 105

first-grade classes and 57 third-grade classes, respectively. McDonald and Elias (1976, p. 534) reported such multiple Rs as .60 between four observation system variables and adjusted pupil reading achievement in 33 second-grade classes. Results of this kind suggest that, as teacher behavior variables are considered in combination, their relationship to pupil achievement can become substantial.

But we need experiments for greater confidence that a difference in teacher behavior causes a difference in student achievement or attitude. My co-workers and I (Program on Teaching Effectiveness, SCRDT, 1976) have conducted an experiment on teaching to determine, in a way more complex than had previously been tried, whether manipulations of teacher behavior, applied to pupils randomly assigned to the various treatments, would influence achievement. Three teacher behaviors were manipulated at two levels in a factorial design that made possible estimation of the main effect of a single teacher behavior variable and also the additive and interactive effects of two or more behaviors varied together.

In that experiment, the manipulations of teacher behavior did make a difference in adjusted class achievement. Furthermore, the effects of the three manipulated dimensions of teacher behavior were additive. When the teacher's "structuring" was at a high level as against low, the adjusted mean achievement on a test, consisting of items whose answers were available only from the teacher and not also from the text, increased by .31 points. When teachers did more low-level questioning as against high-level questioning, the resulting increase was .78. When teachers did high reacting as against low reacting, the increase was .37. The sum of these main effects is 1.46. When all three of these dimensions of teacher

behavior were manipulated together, the difference between the best combination of the three dimensions and the worst was 1.57 points, only slightly greater than the sum of the three one-dimensional differences. In short, the differences produced when three dimensions of teacher behavior were manipulated were additive. And, since these results were obtained in a true experiment, it seems safe to infer that the differences in teacher behavior actually caused the differences in achievement. Since there are dozens of dimensions of teacher behavior, the best combination of teacher behaviors ought to make a substantial difference in student achievement.

The Power of Pre-Instructional Variables

Finally we turn to the power of pre-instructional variables. In every study of teaching, so far as I know, such variables have accounted for more of the variance in pupil achievement and attitude than the teacher or teaching variables have accounted for. Typically, a measure of the pupils' scholastic aptitude or prior achievement, obtained prior to the beginning of instruction, correlates much more highly than teacher behavior variables with measures of pupil achievement obtained at the end of the teaching period. This result prevails whether the teaching lasts for fifteen minutes or a whole school year. The correlations of pre-instructional variables with post-instructional achievement may be as high as .7 to .9.

Investigators have regarded this fact as a reason for despair about the importance of variables in teaching. Too little variance in achievement remains to be accounted for. The total importance of variables in teaching seems doomed to remain what Heath and Nielson called "inherently

trivial" (1974, p. 481).

But it may well be that teaching-behavior variables have been forced into an unfair comparison. The pre-instructional variables against which they have been pitted represent the outcome of years of prior experience and development. During these years, pupil ability and achievement have been influenced not only by genetic variables of much debated importance. They have also been influenced by dozens of powerful "teachers" in the form of parents, friends, neighborhoods, mass media, and even persons called teachers. So the pupil who arrives at the classroom door in September has already been much "taught."

The influences of those years of prior "teaching," by "teachers" who have had much more prolonged and intimate opportunities to exert their influence, must be regarded as prior links in a long chain of cause-effect relationships. In any model of those causal relationships, the last box representing "pupil achievement" becomes the first box representing "pre-instructional variables" in the next round of teacher influence on the pupil. Because the effects are cumulative, with each "teacher" adding to the effects of the preceding "teachers," the last teacher in the chain cannot be expected to have an effect nearly equal to the total effect of all his or her predecessors. Yet it is that last teacher in the chain whose effect is studied in any investigation of the importance of teaching variables. Viewed in this way, the last teacher will inevitably have a "trivial" effect.

But if the variables characterizing the last teacher are trivial, are not those descriptive of the next-to-last teacher also trivial? And those of the teacher before that? Such reasoning leads to the conclusion

that the pupil achievement that we are trying to understand and improve has resulted from a long chain of trivial influences. But the pupil achievement, and its variance, is obviously nontrivial--it is the knowledge, understanding, and attitudes on which our civilization rests. Hence, something must be wrong with the methods of analysis that lead us to conclude that teaching variables are trivial. We need alternative kinds of reasoning.

One alternative is to develop ways of cumulating the effects of teacher variables over successive teachers. Suppose the teachers of three-year olds produce 10% of the total variance in achievement among their pupils, and the teachers of four-year olds produce another, partially independent, 10%, and the kindergarten teachers produce another, partially independent, 10%, and so on, through the school grades. By the time the pupils' school careers have been completed, the total effects of all their teachers, in terms of the variance in their achievement attributable to their teachers, will have become nontrivial. Yet the numbers used in this illustration are of the order of magnitude that characterizes the results of present-day studies of teaching (cf. the multiple correlations cited above). Accordingly, if methods become available for estimating the total importance of teacher variables over all of a pupils' teachers, we can expect that importance to be far less overshadowed by that of preinstructional variables. Teacher variables will no longer be deprived of credit for having produced some of the preinstructional variance that now makes the variance attributable to teachers look so weak.

Summary

The apparently discouraging record of research on teaching has resulted in part from inadequate ways of accumulating evidence, characterized by attaching excessive importance to the statistical significance of single studies. Because those studies have typically been based on small groups of teachers and the expected correlation of single teacher variables with pupil achievement is low, the typical study has been conducive to making the error of considering false null hypotheses to be true.

The chi-square model for estimating the significance of combined results makes possible a more valid accumulation of the evidence. When applied to several clusters of studies, that model yielded estimates indicating that certain teacher variables do indeed correlate significantly with adjusted pupil achievement or attitude. Previous reviews are thus seen to have resulted in unjustifiably equivocal or discouraging conclusions. The approach also implies that a number of small-sample independent studies have greater veridical value than a single large-sample study. Especially when performed by different investigators, the combined results of several small studies can equal the larger one in statistical power and gain the additional advantage of generalizability across investigators.

Certain additional considerations also permit a less pessimistic view of the promise of research on teaching. Interaction effects, including interactions of teaching variables with pupil "aptitude," may indeed

occur. But research on teaching, and the behavioral sciences in general, are not in principle in any worse situation than the natural sciences in this respect. Single laws (i.e., main effects) hold strongly only in laboratories where other variables can be controlled. In the real world, whether it is the world of the falling leaf or that of the classroom, such laws have explanatory power only when used in two's, three's, or larger combinations.

Further, it is true that the correlations that research on teaching has typically yielded are low. But such correlations can nonetheless be important in terms of the numbers of pupils who could be affected by many teachers over many years at almost no additional cost for the implied better ways of teaching. And the correlations become much more substantial when based on several teacher variables, as in computing multiple correlations. Further, in at least one experiment, the effects of three independent teacher variables were found to be additive, with the implication that the total causal influence of a number of teacher variables could be substantial.

Finally, the power of preinstructional pupil characteristics in determining pupil achievement was considered. Comparing the power of teacher variables to that power may lead to inappropriately low estimates of the importance of teacher variables. The preinstructional variables represent the outcome of years of such prior influences as parents, mass media, and even persons called teachers. More appropriate estimates of

the importance of teacher variables may be made possible by developing methods for summing the variances attributable to clusters of teachers in successive grade levels as pupils move through their school careers.

In sum, the long prevalent disparagement of the results of the search for relationships between teacher behavior and student achievement seems to need reconsideration. Properly viewed, results have been far from entirely negative, inconsistent, or inconsequential. The intuitive judgment of millions of parents and hundreds of investigators--that teacher variables of certain kinds do make a difference in pupil achievement and attitude--has been sufficiently borne out to justify four cheers for research on teaching.

Notes

1. As is noted in the text, Dale Schunk assisted me in applying the chi-square model to estimating the significance of combined results. My work on this paper was conducted in the Program on Teaching Effectiveness of the Stanford Center for Research and Development in Teaching, which is supported in part by the National Institute of Education, Department of Health, Education, and Welfare (Contract No. NE-C-00-3-0061). The opinions expressed in this paper do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. Additional support was provided by a fellowship from the John Simon Guggenheim Memorial Foundation for the academic year 1976-1977. An earlier version of this paper was presented as an invited address to the Special Interest Group on Research in Mathematics Education at the Annual Meeting of the American Educational Research Association, San Francisco, April 20, 1976, under the title "Issues in the Field of Research on Teaching."
2. For the three cited clusters of studies, tables similar to Tables 1 and 2 are available in N. L. Gage, The Scientific Basis of the Art of Teaching, New York: Teachers College Press, 1978. The argument of the present paper is extended in that volume.

Table 1

Teacher Indirectness and Pupil Achievement, Field Surveys (Based on Dunkin and Biddle, 1974, p. 115)

Study	Duration of Teaching	Grade Level	No. of Teachers	Definition of Independent Variable	Definition of Dependent Variable	Result	Probability of Result	χ^2
Flanders (1970)	2 semesters	2	15	Factor I (i/i+d)	"Portions of a nationally standardized achievement test covering language and number skills were used to form a composite score" (p. 393)	r=-.073	.605	1.0051
Flanders (1970)	2 weeks	4	16	Factor I (i/i+d)	Achievement test covering a two-week unit in social studies	r=.308	.123	4.1911
Flanders (1970)	2 semesters	6	30	Factor I (i/i+d)	Metropolitan Achievement Test mean score	4=.224	.142	3.9039
Flanders (1970)	2 weeks	7	15	Factor I (i/i+d)	Special social studies unit on New Zealand	r=.481	.035	6.7048
Flanders (1970)	2 weeks	8	16	Factor I (i/i+d)	Special mathematics unit dealing with applications of algebraic and geometric formulas	r=.428	.049	6.0319
Cook (1967)	2 semesters	10	8	Flanders I/D ratio (pp. 26, 27)	Watson-Glaser Critical Thinking Appraisal	χ^2=13.33, df=7	.067	5.4061
					Processes of Science Test	χ^2=9.55, df=7	.217	3.0557
					Comprehensive Final Test	χ^2=9.55, df=7	.217	3.0557

Table 1 (cont.)

Study	Duration of Teaching	Grade Level	No. of Teachers	Definition of Independent Variable	Definition of Dependent Variable	Result	Probability of Result	χ^2
Furst (1967)	4 one-hour lessons	10, 12	15	40-minute achievement test on international economic problems (Same as Bellack,et al.,1966). Three groups consisting of 3 high-achieving, 8 average, and 4 low-achieving classes (See Footnote 1)	Composite of three Flanders variables:(a) ratio of extended indirect teacher influence to extended direct teacher influence;(b) revised i/d in response to pupil talk;(c) extended student talk	l=3.895 df=2, 12	.024	7.4594
Medley and Mitzel (1959)	2 semesters	3-6	49	"Emotional Climate" as measured by OScAR (p. 241)	Four subtests of the California Reading Test (elementary): word form, word recognition, meaning of similarities, and interpretation of meanings	r=.20	.081	5.0266
Powell (1968)	2 semesters	3	9	See Footnote 2	SRA Achievement Series: Composite	F=5.85, df=1,164	.007	9.9237
					SRA Reading Total	F=1.30, df=1,164	.127	4.1271
					SRA Arithmetic Total	F=10.68, df=1,164	.001	13.8155

[1]Note that, in this study, student achievement became the independent variable.

[2]The Flanders system of interaction analysis was...used. From the observational data collected, percentage matrixes were built from which scores for each teacher were extracted to form multivariate composite scores. These were: i/d Ratio; the ratio of motivating to controlling teacher behaviors; i/d Rows 8 and 9, which is similar to the i/d Ratio except that it only includes teacher response to pupil talk; extended i/d Ratio, which is similar to those above except that it examines extended motivation or control; 9-3 Cell, which represents the number of times the teacher accepted and used pupils' ideas, as shown on the matrix; 9-9 Cell representing extended pupil talk about their own ideas; and the 3-3 Cell, which represents teacher extension of pupils' ideas for some time.

Table 1 (cont.)

Study	Duration of Teaching	Grade Level	No. of Teachers	Definition of Independent Variable	Definition of Dependent Variable	Result	Probability of Result	χ^2
Snider (1966)	2 semesters	12	17	Flanders I/I+D Ratio: Five most direct teachers were compared with the five least direct teachers		Mann-Whitney		
					New York Regents Physics Examination	U=18	.155	3.7287
					Cooperative Physics Test	U=13	.500	1.3863
					Test of Understanding Science-I	U=12	.500	1.3863
					Test of Understanding Science (total)	U=14	.421	1.7302
Weber (1968)	3 years	4	26	Classification of teacher on the basis of multivariate composite scores in Flanders Interaction Analysis Categories: 4 groups of pupils (n=45 each) formed as follows: (First 3 years had indirect teacher; 4th year indirect teacher; First 3 years indirect teacher, 4th year direct teacher; First 3 years direct teacher, 4th year indirect teacher; First 3 years direct teacher, 4th year direct teacher)	Torrance Test of Creative Thinking, verbal fluency score	For comparison of students who in 1st 3 years had indirect teachers vs. direct teachers, F=10.58, df=1,176	.001	13.8155
Thompson and Bowers (1968)	2 semesters	4	15	Teacher supportiveness (similar to "emotional climate" of the Medley-Mitzel OScAR-2e)	Stanford Achievement Test Vocabulary score	F < 1.0 df=1, 13	-	-
					Stanford Achievement Test Social Studies score	F < 2.0 df=1, 13	-	-

Table 1 (cont.)

Study	Duration of Teaching	Grade Level	No. of Teachers	Definition of Independent Variable	Definition of Dependent Variable	Result	Probability of Result	χ^2
Torrance and Parent (1966)	2 semesters	7-12	10	Flanders I/D Ratio	Sequential tests of educational progress, mathematics gain score	rho=.32	.184	3.3856
Allen (1970)	2 semesters	1	18	Record of verbal behavior (developed from OScAR 4V warmth score)	Number attainment test	(see Rosenshine, 1971, p.79, since not presented in Allen's report)	.83	.3727
					Number extension test	ditto	.83	.3727
					Number concepts test	ditto	.79	.4714
Soar (1966)	2 semesters	3-6	55	Rotated Factor 8: Indirectness vs. silence and confusion	Iowa Test of Basic Skills:			
					Vocabulary	r=.068	.322	2.2664
					Reading	r=.021	.443	1.6284
					Arithmetic Concepts	r=.034	.408	1.7930
					Arithmetic Problems	r=.083	.309	2.3488
					Arithmetic Total	r=.081	.314	2.3167
Soar et al. (1971)	6 months	kindergarten	35	Factor I: Free choice vs. structured learning in groups	Lee-Clark Reading Readiness Test, 1/3 of original test	r=.00	.500	1.3863
	6 months	1	20	Same as for kindergarten	Metropolitan Readiness Tests, 1/3 original length	r=.30	.099	4.6253

Table 1 (cont.)

Study	Duration of Teaching	Grade Level	No. of Teachers	Definition of Independent Variable	Definition of Dependent Variable	Result	Probability of Result	χ^2
Hunter (1968)	2 semesters	ages 8-14	11	Withall Scale: I/D Ratio similar to Flanders	Wide Range Achievement Test: Median score on reading, spelling, and arithmetic	r=.62	.021	7.7265
LaShier (1967)	6 weeks	8	10	Flanders I/D Ratio	Biological Sciences Curriculum Study Unit Test	tau=.60	.008	9.6566
Pinney (1969)	Two 45-minute lessons	8-9	32[1]	Percentage of times teacher followed a pupil response with two or more reinforcing statements (equivalent to "extended indirectness" in Flanders)	Achievement tests for social studies and English	F=4.2 df=1, 30	.170	3.5439

[1]16 most and 16 least effective (as defined by adjusted student achievement) out of 56 teachers

Table 2. Teacher Indirectness and Student Achievement: Field Surveys Alternative Findings for Three Methods of Combining Results Shown in Table 1

Study	Most Significant Chi Squares	Mean Chi Squares	Least Significant Chi Squares
Flanders (1970) Grade 2	1.0051	1.0051	1.0051
Flanders (1970) Grade 4	4.1911	4.1911	4.1911
Flanders (1970) Grade 6	3.9039	3.9039	3.9039
Flanders (1970) Grade 7	6.7048	6.7048	6.7048
Flanders (1970) Grade 8	6.0319	6.0319	6.0319
Cook (1967)	5.4061	3.8391	3.0557
Furst (1967)	7.4594	7.4594	7.4594
Medley & Mitzel (1959)	5.0266	5.0266	5.0266
Powell (1968)[a]	-	-	-
Snider (1966)	3.7287	2.0578	1.3863
Soar (1966)	2.3488	2.0706	1.6284
Soar, et al. (1971)	4.6253	3.0058	1.3863
Hunter (1968)	7.7265	7.7265	7.7265
LaShier (1965)	9.6566	9.6566	9.6566
Pinney (1969)	3.5439	3.5439	3.5439
Weber (1968)[a]	-	-	-
Thompson & Bowers (1968)[a]	-	-	-
Torrance & Parent (1966)	3.3856	3.3856	3.3856
Allen (1970)	.4714	.4056	.3727
Sum of Chi Squares	75.2157	70.0143	66.4648
Probability of Combined Result*	$p < .001$	$p < .001$	$p < .001$

*$\chi^2_{(df=36,\ p=.001)}=67.985$

[a]Data omitted because the significance test was based on the pupil as the unit of analysis.

Table 3

Frequency of Smoking in
Lung Cancer Patients and Controls

(Based on data in Lilienfeld, et al., 1967, p. 78)

		Lung Cancer		
		Patients	Controls	Total
Smoking	Yes	499	462	961
	No	19	56	75
	Total	518	518	1036

$\phi = .14$

References

American Educational Research Association, Committee on the Criteria of Teacher Effectiveness. Report of . . . . Review of Educational Research, 1952, 22, 238-263.

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J. How effective is schooling? A critical review and synthesis of research findings. Santa Monica, Cal.: Rand Corp., 1971.

Baker, J. P., & Crist, J. Teacher expectancies: A review of the literature. In J. D. Elashoff & R. E. Snow (Eds.), Pygmalion reconsidered. Chicago: National Society for the Study of Education, 1971. Pp. 48-64.

Bonarth, J. D., & Roberts, R. R., Jr. Signifying significant significance. American Psychologist, 1972, 27, 774-775.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 171-246.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. American Psychologist, 1975, 30, 116-127.

Do teachers make a difference? A report on recent research on pupil achievement. Washington, D. C.: Bureau of Educational Personnel Development, Office of Education, 1970.

Doyle, W. Paradigms in teacher effectiveness research. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April 1975.

Dubin, R., & Taveggia, T. C. The teaching-learning paradox: A comparative analysis of college teaching methods. Eugene, Ore.: Center for the Advanced Study of Educational Administration, 1968.

Dunkin, M. J., & Biddle, B. J. The study of teaching. New York: Holt, Rinehart & Winston, 1974.

Getzels, J. W., & Jackson, P. W. The teacher's personality and characteristics. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 506-582.

Glass, G. V. The wisdom of scientific inquiry on education. Journal of Research in Science Teaching, 1972, 9, 3-18.

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5 (10), 3-8.

Gordon, N. H., Loveland, E. H., & Cureton, E. E. An extended table of chi-square for two degrees of freedom, for use in combining probabilities from independent samples. Psychometrika, 1952, 17, 311-316.

Heath, R. W., & Nielson, M. A. The research basis for performance-based teacher education. Review of Educational Research, 1974, 44, 463-483.

Holton, G. On the role of themata in scientific thought. Science, 1975, 188, 328-338.

Hunt, K. Do we really need more replications? Psychological Reports, 1975, 36, 587-593.

Jones, L. V., & Fiske, D. W. Models for testing the signficance of combined results. Psychological Bulletin, 1953, 50, 375-382.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Lilienfeld, A. M., Pedersen, E., & Dowd, J. E. Cancer epidemiology: Methods of study. Baltimore: Johns Hopkins Press, 1967.

McDonald, F. J., & Elias, P. J. The effects of teaching performances on pupil learning. Beginning teacher evaluations study: Phase II. Final report: Vol. I. Princeton, N. J.: Educational Testing Service, 1976.

McKeachie, W. J., & Kulik, J. A. Effective college teaching. In F. N. Kerlinger (Ed.), Review of Research in Education, Vol. 3. Itasca, Ill.: F. E. Peacock, 1975. Pp. 165-209.

Mosteller, F., & Moynihan, D. On equality of educational opportunity. New York: Random House, 1972.

Program on Teaching Effectiveness, SCRDT. A factorially designed experiment on teacher structuring, soliciting, and reacting. Stanford, Cal.: Stanford Center for Research and Development in Teaching, 1976.

Rosenshine, B. Teaching behaviours and student achievement. Slough, England: National Foundation for Educational Research in England and Wales, 1971.

Rosenshine, B., & Furst, N. F. Research on teacher performance criteria. In B. O. Smith (Ed.), Research in teacher education: A symposium. Englewood Cliffs, N. J.: Prentice-Hall, 1971. Pp. 37-72.

Stallings, J. A., & Kaskowitz, D. H. Follow through classroom observation evaluation: 1972-1973. Menlo Park, Cal.: Stanford Research Institute, 1974.

Stephens, J. M. The process of schooling. New York: Holt, Rinehart & Winston, 1967.

Sterling, T. C. Publication decisions and their possible effects on inference drawn from tests of significance--or vice-versa. Journal of the American Statistical Association, 1959, 54, 30-34.

Wallen, N. E., & Travers, R. M. W. Analysis and investigation of teaching methods. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 448-505.

Withall, J., & Lewis, W. W. Social interaction in the classroom. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 683-714.

Wright, C. J., & Nuthall, G. Relationships between teacher behaviors and pupil achievement in three experimental elementary science lessons. American Educational Research Journal, 1970, 7, 477-491.